АБУМАТРАН
ABUMATRAN

The Abu-MaTran project: tools for teaching machine translation

Víctor M. Sánchez-Cartagena
Prompsit Language Engineering, S.L.

DCU prompsit Universitat d'Alacant Universidad de Alicante IEA ILSP

Outline

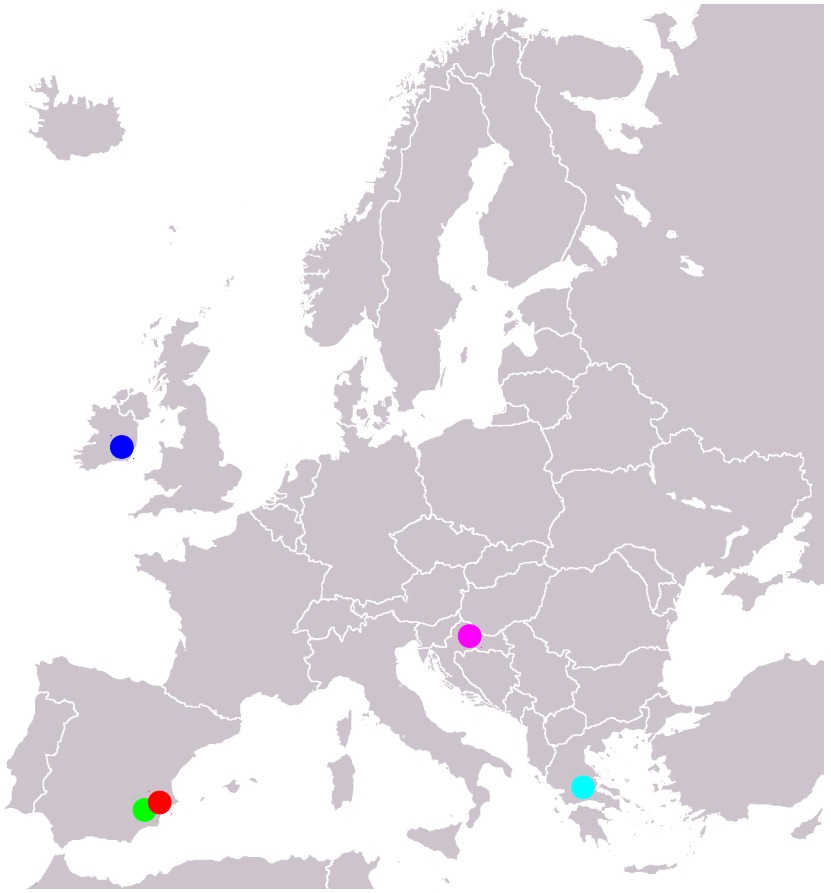
1) The Abu-MaTran project in a nutshell
2) Acquisition of parallel data from the web
   - How a web crawler works
   - Web crawling in the Abu-MaTran project
   - Hands-on session: Bicrawler
3) Building statistical machine translation (SMT) systems
   - Introduction to SMT
   - SMT systems released in the Abu-MaTran project
   - Hands-on session: MTradumàtica

The Abu-MaTran project in a nutshell

Abu-MaTran in a nutshell

- Project type: Marie Curie IAPP (Industry-Academia Partnerships and Pathways)
  - core activity: transfer of knowledge
  - by means of secondments: put in contact academic and industrial partners
- Duration: 48 months (from January 2013): it is about to end

Partners



- Dublin City University (Ireland)
- Prompsit Language Engineering (Spain)
- University of Alicante (Spain)
- University of Zagreb (Croatia)
- Institute for Language and Speech Processing (Greece)

Abu-MaTran in a nutshell

- Enhance industry-academia cooperation to tackle multilinguality
- Increase low industrial adoption of machine translation
- Transfer back to academia the know-how of industry to make research products more robust
- Resources produced to be released as free/open-source software
- Focus on Croatian: language of new EU member state
- Emphasis on dissemination

Some results (I)

- Open-source software released:
  - 2 web crawlers
  - Tool for getting corpora from Twitter
  - Tool for inferring shallow-transfer rules from small parallel corpora
  - Tool for adding entries to RBMT monolingual dictionaries
- Corpora released:
  - General-domain monolingual corpora for Croatian, Serbian, Bosnian, Catalan and Finnish
  - Tweets monolingual corpora for Croatian, Serbian and Slovene
  - General-domain parallel corpora for English-to Croatian, Serbian, Bosnian and Finnish
  - Tourism parallel corpora for English-Croatian
  - ...

Some results (II)

- MT systems created:
  - RBMT: Serbian-Croatian
  - SMT: domain adaptation and linguistic resources:
    - Tourism domain English-Croatian
    - General domain English-Croatian
    - Tourism domain English-Greek
- Participation in shared tasks
  - Winning systems in WMT 2014,2015,2016
  - Winning systems TweetMT 2015

Some results (III)

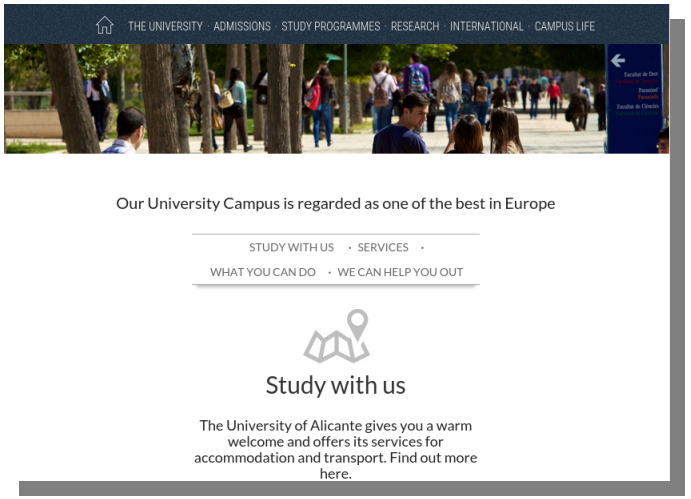
- Organization of Spanish Linguistics Olympiad 2014-2015-2016
- Workshop organization:
  - 2014, DCU: Software management for researchers
  - 2014-2015, Zagreb: data creation for Croatian RBMT
  - 2014, Reykjavik: free/open-source RBMT linguistic resources
  - 2016, DCU: Hybrid machine translation
  - 2016, DCU: Tools for linguists

Acquisition of parallel data from the web

1)How a web crawler works
2)Web crawling in the Abu-MaTran project
3)Hands-on session: Bicrawler

How a web crawler works

- How can we turn a multilingual website ...



- … into a parallel corpus ready for SMT?

Our University Campus is regarded as one the best in Europe	La Universidad puede presumir de tener uno de los mejores campus europeos
Study with us	¿Vienes?

How a web crawler works

1) Download web pages
2) Extract text and remove HTML tags
3) Detect language of documents
4) Identify documents that are mutual translation **(most difficult part)**
5) Extract parallel sentences from each document pair

How a web crawler works

1)Download web pages

- The most time-consuming part: downloading a big website can take days!
- From the main page (e.g. www.ua.es), hyperlinks are followed in order to get new documents
- From new documents, hyperlinks are followed in order to get more documents, and so on…
- It is very important to follow the rules in **robots.txt**

How a web crawler works

2)Extract text and remove HTML tags

- HTML tags need to be stored: they are needed in subsequent steps
- Text is split into **paragraphs**

```
<div class="row">
<div class="col-md-12">
<h2 class="subSeccionIcono"
id="vienes"><img
src="https://web.ua.es/secciones-
ua/images/acceso/estudia/vida-
universitaria/icono1.jpg" /> Study with
us</h2>
<h3 class="subtituloIcono">The University
of Alicante gives you a warm welcome and
offers its services for accommodation and
transport. Find out more here.</h3>
```

Study with us

The University of Alicante gives you a warm welcome and offers its services for accommodation and transport. Find out more here.

How a web crawler works

3)Detect language of documents

Study with us

The University of Alicante gives you a warm welcome and offers its services for accommodation and transport. Find out more here.

English

¿Vienes?

La Universidad de Alicante te acoge con toda clase de facilidades para el alojamiento o el transporte. Conócelas aquí.

Spanish

How a web crawler works

4) Identify documents that are mutual translation

- The most difficult part
- There is a shared task at WMT conference
- Clues that help us to identify pairs of documents:
  - URL: e.g. https://web.ua.es/en/university-life.html and https://web.ua.es/es/university-life.html
  - Images
  - Numbers
  - Named entities
  - HTML structure/layout
  - Links
  - Similarity after being translated with some bilingual resource: finding parallel resources is difficult for some language pairs!

How a web crawler works

5) Extract parallel sentences from each document pair

- Don’t join sentences from different paragraphs

Study with us	¿Vienes?
The University of Alicante gives you a warm welcome and offers its services for accommodation and transport.	La Universidad de Alicante te acoge con toda clase de facilidades para el alojamiento o el transporte.
Find out more here.	Conócelas aquí.

How a web crawler works

5) Extract parallel sentences from each document pair

- Don’t join sentences from different paragraphs

Language promoter and specialist in language planning**.** Professionals in this area offer services associated with standardisation, linguistic planning and language promotion. Professionals work with language users and study their linguistic behaviour.

Dinamizador lingüístico y especialista en planificación lingüística**:** se trata de un profesional que presta servicios vinculados a la normalización, la planificación lingüística y la promoción de una lengua. La materia de trabajo de este profesional son los usuarios y sus comportamientos lingüísticos.

Language promoter and specialist in language planning**.** Professionals in this area offer services associated with standardisation, linguistic planning and language promotion	Dinamizador lingüístico y especialista en planificación lingüística**:** se trata de un profesional que presta servicios vinculados a la normalización, la planificación lingüística y la promoción de una lengua.
Professionals work with language users and study their linguistic behaviour.	La materia de trabajo de este profesional son los usuarios y sus comportamientos lingüísticos.

Crawling tools developed

- Bitextor: http://bitextor.sourceforge.net/
  - Developed by Prompsit Language Engineering and University of Alicante
  - Produces a parallel corpus from a mutilingual web site
  - Needs bilingual lexicon
  - Document alignment by means of automatic classifier
- ILSP-FC: http://nlp.ilsp.gr/redmine/projects/ilsp-fc
  - Developed by ILSP (Greece)
  - Can be used to produce monolingual or parallel corpora, from multiple websites and even a list of terms
  - Does not need any bilingual resource
  - Document alignment by means of heuristics

Monolingual corpora

- Important resource for SMT: building language models
- From Internet top-level domains:
  - `.hr` (Croatian; 1340M toks.), `.bs` (Bosnian; 288M toks.), `.sr` (Serbian; 557M toks.) → English-Croatian tourism SMT
  - `.fi` (Finnish; 1700M toks.) → WMT 2015 good results
  - `.cat` (Catalan; 779M toks.)
- From Twitter:
  - With our tool TweetCaT: 236M toks. for Serbian/Croatian, 38M toks. for Slovene

Parallel corpora

- Even more important resource for SMT: more difficult to find
- From Internet top-level domains, with Bitextor+Spiderling:
  - `.sl` (Slovene-English; 37M toks.)
  - `.sr` (Serbian-English; 27M toks.)
  - `.hr` (Croatian-English; 71M toks.)→ English-Croatian SMT
  - `.fi` (Finnish-English; 100M toks.) →WMT 2015 good results
- From lists of websites, with ILSP-FC:
  - Croatian tourism websites (Croatian-English; 146k segments) → English-Croatian tourism SMT
  - Greek tourism/culture websites (Greek-English; 4M toks.) → English-Greek tourism SMT

Bicrawler

- Web-based service for extracting parallel corpora from multilingual websites
- Makes acquisition of parallel data available to everyone
- Developed by Prompsit Language Engineering
- Built upon the two open-source web crawlers released during the project: Bitextor and ILSP-FC
- Added an additional cleaning layer to remove possible errors introduced by the crawling tools
- Free use, but limited in terms of crawling time
- Unlimited (premium) version will be available soon

Hands-on session

Download instructions from http://abumatran.eu/dcu-nov-2016-guide.pdf

Building SMT systems

1)Introduction to SMT
2)SMT systems deployed in the Abu-MaTran project
3)Hands-on session: MTradumàtica

Statistical machine translation

- **Translation**: TL sentence with highest probability according to a combination of statistical models
- Translation hypotheses are built by splitting the SL sentence in segments and concatenating (not necessarily in the same order) their translations according to a phrase translation model
- Example: *the small houses*

the	*el*	0.5
the	*las*	0.2
the small	*el*	0.05
small houses	*casas pequeñas*	0.7
small	*medianas*	0.1
houses	*hogar*	0.3

el casas pequeñas	0.35
el hogar	0.015
las casas pequeñas	0.14
el medianas hogares	0.015

Different models

- **Phrase translation model** in both directions
- Language model of the target language (TL)
- Word penalty
- Phrase penalty
- Reordering model
- ...

Phrase translation model

- Phrase table
  - Multi-word probabilistic bilingual dictionary (in both directions) with variable-length segments

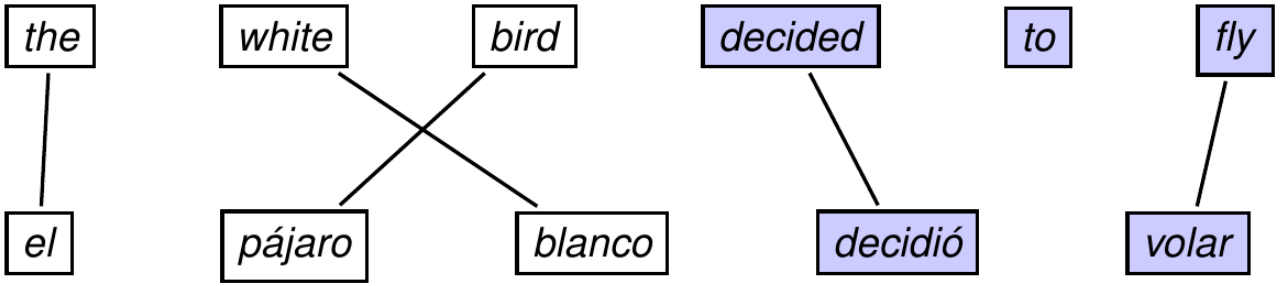
Source (s)	Target (t)	$p(s\|t)$	$p(t\|s)$
...	...	...	...
here are the dates for	voilà les dates de	1.00	1.00
here are the dates	voilà les dates	1.00	1.00
here are the	voici donc les	0.33	0.50
here are the	voilà les	0.04	0.50
...	...	...	...

Phrase translation model

Obtained from a **parallel corpus**

1)Compute word alignments

2)Extract bilingual phrases from the word alignments



3)Compute translation probabilities

$$p(s|t) = \frac{\text{count}(s \leftrightarrow t)}{\text{count}(t)}; \quad p(t|s) = \frac{\text{count}(s \leftrightarrow t)}{\text{count}(s)}$$

Target language model

- It allows us to measure how likely (fluent) a TL sentence is, how “good” it is that sentence in the TL
- Usually: statistical model based on n-grams (segments of n words)

$$
\begin{aligned}
&p(\text{The potential of machine translation is clear}) = \\
&\quad p(\text{The}) \times p(\text{potential}|\text{The}) \times p(\text{of}|\text{The potential}) \times \\
&\quad p(\text{machine}|\text{potential of}) \times p(\text{translation}|\text{of machine}) \times \\
&\quad p(\text{is}|\text{machine translation}) \times p(\text{clear}|\text{translation is})
\end{aligned}
$$

- Easily obtained from large TL (monolingual) texts:

$$p(\text{house}|\text{the red}) = \frac{\text{count}(\text{the red house})}{\text{count}(\text{the red} *)}$$

Other models

- Word penalty: number of words in the target translation
  - The language model likes short sentences (less n-grams to score)
  - Used to avoid producing very short translations
- Phrase penalty: number of bilingual phrases used to produce the target
  - Used to promote the use of long phrases (fewer phrases)
- Reordering model: how likely is to change the order of a phrase when assembling the translation hypothesis.

Parameter tuning

- Not all models are equally important
- Probability of a translation hypothesis:

$$p(\text{target}|\text{source}) \propto \lambda_1 h_1(\cdot) + \lambda_2 h_2(\cdot) + \cdots + \lambda_{14} h_{14}(\cdot)$$

- $h_i(.)$: prob of hypothesis according to model; λ_i : weight of model h_i
- Tuning: starting with random values for the weights λ_i, find the set of values that maximises translation quality
  - From a (small) development parallel corpus
  - Its SL side is translated, compared to the TL side and weights are updated to obtain a more accurate translation
  - The process is repeated iteratively

Abu-MaTran SMT systems

- English-Croatian: generic and tourism domain
- WMT 2014: English-French
- WMT 2015: English-Finnish
- WMT 2016*: English-Finnish (NMT)
- English-Greek: tourism/culture domain

English-Croatian SMT systems (I)

- Objective: build a generic (news) and a tourism-oriented SMT system
- Challenges:
  - Available parallel data is generally noisy or out-of-domain:
    - DGT and JCR (law)
    - OpenSubtitles (movie subtitles, noisy)
    - TED Talks (spoken language)
    - SETimes (news)
  - Croatian is a highly inflected language: data sparseness

English-Croatian SMT systems (II)

- Obtain additional parallel data:
  - Crawl `.hr` TLD and tourism web sites
  - Translate Serbian side of English-Serbian parallel data
- Select most appropriate sentences from out-of-domain data using LM perplexity difference (*data selection*)
- Use factored translation models for English→Croatian:

Predsjednik	Karzai	ne	želi	nikakvu	stranoj	kontrolu
N-M-SG-NOM	PN	ADV	VB	DT	ADJ-F-SG-LOC	N-F-SG-ACC ↓
Predsjednik	Karzai	ne	želi	nikakvu	stranu	kontrolu
N-M-SG-NOM	PN	ADV	VB	DT	ADJ-F-SG-ACC	N-F-SG-ACC ↑

- Results:
  - General-domain: outperforms Google Translate
  - Tourism system: outperforms general-domain system when translating tourism websites

WMT participation (I)

- **W**orkshop on **S**tatistical Machine **T**ranslation
  - Annual competition: build the best MT system for the news domain
  - Constrained: from the resources provided
  - Unconstrained: from any resource you can find
- WMT 2014: English → French
  - Data selection: use subset of training data likely to belong to news domain according to LM perplexity
  - Ranked 1st constrained

WMT participation (II)

- WMT 2015: English-Finnish
  - Morphological segmentation on Finnish with `Omorfi` lexicon-based tool in order to deal with data sparseness:
    - Splits compound words
    - Splits simple words in lemma + morphological affixes
  - Ranked 1st constrained, 2nd unconstrained (+ crawled data)

Finnish text	haluaisimme , että oppisimme tästä yhden perusasian
Segmented	halua→ ←isi→ ←mme , että opp→ ←isi→ ←mme tästä yhde→ ←n perus→ `(basic)` ←asia→ `(issue)` ←n *`(case marker)`*

- WMT 2016: English →Finnish
  - Neural MT + morphological segmentation
  - Ranked 1st constrained

English-Greek tourism SMT systems

- Previously built SMT systems followed data selection as the main *domain adaptation* method
- Domain adaptation: method to combine in-domain and out-of-domain data so as to maximize translation quality
- We experimented with different domain adaptation methods in the literature and picked the best ones for our domain:
  - English→Greek: one LM for each domain
  - Greek → English: data selection + different LMs
- Our SMT systems are available at http//translator.abumatran.eu

АБУМАТРАН
ABUMATRAN

Abu-MaTran Translator

Type or paste a text in the *Original text* field, then press *Translate*. The first 1000 characters will be translated.

Direction English – Croatian Translate

Original text

Our translator is working pretty well.

Translated text

Naš prevoditelj radi prilično dobro.

More info about Abu-MaTran

Mtradumàtica (I)

- Web interface for Moses
- Developed by Prompsit Language Engineering for Universitat Autònoma de Barcelona
- Released as open-source software
- Allows you to easily experiment with SMT:
  - Manage files and corpora
  - Train LMs and SMT systems
  - Tune systems
  - Translate text
  - Inspect phrase table and language model

Mtradumàtica (II)

- Currently you cannot:
  - Apply data selection
  - Merge systems with domain adaptation methods
  - Evaluate systems with automatic metrics

- Useful tool for making students understand how SMT works

Hands-on session

Download instructions from http://abumatran.eu/dcu-nov-2016-guide.pdf

АБУМАТРАН
ABUMATRAN

Thank you for your attention

The Abu-MaTran project

DCU prompsit Universitat d'Alacant Universidad de Alicante IEA ILSP